Submitted to the Statistical Science

Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis

Ewan Cameron and Anthony Pettitt

Queensland University of Technology

Abstract. We investigate the utility to contemporary Bayesian studies of recursive, Gauss-Seidel-type pathways to marginal likelihood estimation characterized by reverse logistic regression and the density of states. Through a pair of illustrative, numerical examples (including mixture modeling of the well-known 'galaxy dataset') we highlight both the remarkable diversity of bridging schemes amenable to recursive normalization and the notable efficiency of the resulting pseudo-mixture densities for gauging prior-sensitivity in the model selection context. Our key theoretical contributions show the connection between the nested sampling identity and the density of states. Further, we introduce a novel heuristic ('thermodynamic integration via importance sampling') for qualifying the role of the bridging sequence in marginal likelihood estimation. An efficient pseudo-mixture density scheme for harnessing the information content of otherwise discarded draws in ellipse-based nested sampling is also introduced.

Key words and phrases: Bayes factor, Bayesian model selection, importance sampling, marginal likelihood, Metropolis-coupled Markov Chain Monte Carlo, nested sampling, normalizing constant, path sampling, reverse logistic regression, thermodynamic integration.

1. INTRODUCTION

Though often secondary to parameter inference in the Bayesian paradigm, the normalizing constant, Z, required to establish the posterior, $\pi(\theta|y)$, as a proper probability density,

$$\pi(\theta|y) = \pi(\theta)L(y|\theta)/Z \text{ where } Z = \int \pi(\theta)L(y|\theta)d\theta, \tag{1}$$

for prior, $\pi(\theta)$, and likelihood, $L(y|\theta)$, nevertheless plays a vital role in the domain of Bayesian model selection and model averaging (Kass & Raftery, 1995; Hoeting et al., 1999). Here Z is generally referred to as either the *marginal likelihood* (i.e., the likelihood of the observed data marginalized [averaged] over the prior density) or the *model evidence*. With the latter term though, one risks the impression

Science and Engineering Faculty, School of Mathematical Sciences (Statistical Science), Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia (e-mail: dr.ewan.cameron@gmail.com)

of over-stating the value of this statistic in the case of limited prior knowledge (cf. Gelman et al. 2004, ch. 6). Problematically, few complex statistical problems admit an analytical solution to Equation 1, nor span such low dimensional spaces ($D(\theta) \lesssim$ 5-10) that direct numerical integration presents a viable alternative. With errors (at least in principle) independent of dimension, *Monte Carlo-based integration methods* have thus become the mode of choice for marginal likelihood estimation across a diverse range of scientific disciplines, from evolutionary biology (Xie et al., 2011; Arima & Tardella, 2012; Baele et al., 2012) and cosmology (Mukherjee et al., 2006; Kilbinger et al., 2010) to quantitative finance (Li et al., 2011) and sociology (Caimo & Friel, 2012).

1.1 Monte Carlo-Based Integration Methods

With the bulk of posterior mass most often constrained within a far smaller volume than that of the prior, the simplest marginal likelihood estimators drawing solely from $\pi(\theta|y)$ or $\pi(\theta)$ cannot be relied upon for model selection purposes. In particular, the harmonic mean estimator (HME; Newton & Raftery 1994),

$$\hat{Z}^H = \left[\sum_{i=1}^{n} 1/n/L(y|\theta_i)\right]^{-1} \quad \text{for } \theta_i \sim \pi(\theta|y), \tag{2}$$

suffers from having infinite variance under common modeling conditions, meaning that its convergence towards the true Z as a one-sided α-stable limit law can be *incredibly slow* (Wolpert & Schmidler, 2012). Even when 'robustified' as per Gelfand & Dey (1994) or Raftery et al. (2007), however, the HME remains notably *insensitive* to changes in $\pi(\theta)$, whereas Z itself is characteristically sensitive (Robert & Wraith, 2009; Friel & Wyse, 2012). Though assuredly finite by default the small-sample variance of the prior arithmetic mean estimator (AME),

$$\hat{Z}^A = \sum_{i=1}^{n} L(y|\theta_i)/n \text{ for } \theta_i \sim \pi(\theta), \tag{3}$$

on the other hand, remains notoriously large, with huge sample sizes again necessary to achieve reasonable accuracy (cf. Neal 1999).

A wealth of more sophisticated integration methods have lately been developed for generating reliable estimates of the marginal likelihood (see Robert & Wraith 2009 and Friel & Wyse 2012 for recent reviews). These include: annealed importance sampling (Neal, 2001), bridge sampling (Meng & Wong, 1996), [ordinary] importance sampling (cf. Liu 2001), path sampling/thermodynamic integration (Gelman & Meng, 1998; Lartillot & Philippe, 2006; Friel & Pettitt, 2008; Calderhead & Girolami, 2009), nested sampling (Skilling, 2006; Feroz & Hobson, 2008), nested importance sampling (Chopin & Robert, 2010), reverse logistic regression (Geyer, 1994), sequential Monte Carlo (SMC; Cappé et al. 2004; Del Moral et al. 2006), and the density of states (Habeck, 2012; Tan et al., 2012). A common thread running through all these schemes is the aim for a superior exploration of the relevant model space via 'guided' transitions across a sequence of intermediate distributions bridging their $\pi(\theta)$ and $\pi(\theta|y)$ extremes. (Or, more generally, their $h(\theta)$ and $\pi(\theta|y)$ extremes if a suitable auxiliary/reference density, $h(\theta)$, is available to facilitate the integration; cf. Lefebvre et al. 2010.) However, the

exact bridging paths specified may be quite dissimilar. Nested sampling, for instance, evolves its particles over a sequence of *constrained-likelihood* distributions, $f(\theta) \propto \pi(\theta)I(L(y|\theta) > L_{\text{lim}})$, transitioning from the prior ($L_{\text{lim}} = 0$) through to the (vicinity of the) peak likelihood ($L_{\text{lim}} \approx L_{\text{max}} - \epsilon$); while thermodynamic integration draws progressively (via Markov Chain Monte Carlo [MCMC]; Tierney 1994) from the family of 'power posteriors',

$$\pi_t(\theta|y) \;\propto\; \pi(\theta)L(y|\theta)^t, \tag{4}$$

most explicitly connecting prior, $\pi(\theta)$ $[t = 0]$, to posterior, $\pi(\theta|y)$ $[t = 1]$.

Another key point of comparison between these rival Monte Carlo techniques lies in their choice of identity by which the evidence is ultimately computed. The [geometric] path sampling identity,

$$\log Z = \int_0^1 E_{\pi_t}\{\log L(y|\theta)\}dt, \tag{5}$$

for instance, is shared across both thermodynamic integration and SMC, in addition to its namesake. However, SMC can also be run with the "stepping-stone" solution (cf. Xie et al. 2011),

$$Z = \prod_{j=2}^{m} Z_{t(j)}/Z_{t(j-1)} \text{ where } t(1) = 0 \text{ and } t(m) = 1, \tag{6}$$

with $\{t(j) : j = 1, \ldots, m\}$ indexing an often parametric sequence of ("tempered") bridging densities (either pre-defined or drawn stochastically from some $\pi(t)$; Gelman & Meng 1998), and indeed this is the mode prefered by experienced practitioners (e.g. Del Moral et al. 2006). Yet another identity for computing the marginal likelihood is that of the recursive pathway, characterised by reverse logistic regression (RLR) and the density of states (DOS).

By *recursive* we mean that, algorithmically, the estimator may be run such that the desired Z is obtained through backwards induction of the complete sequence of intermediate normalizing constants, $Z_{t(j)}$, corresponding to the m indexed bridging densities *by supposing these* $Z_{t(j)}$ *to be already known*. That is, a stable solution may be found in a Gauss-Seidel-type manner by starting with a guess of each normalizing constant as input to a convex system of equations for updating these guesses, returning the new output as input to the same equations, and iterating until convergence. In fact, although the RLR and the DOS approaches differ vastly in concept and derivation—the former emerging from considerations of the reweighting mixtures problem in applied statistics (Geyer & Thompson, 1992; Geyer, 1994; Chen & Shao, 1997; Kong et al., 2003) and the latter from computational strategies for free energy estimation in physics/chemistry/biology (Ferrenberg & Swendsen, 1989; Kumar et al., 1992; Shirts & Chodera, 2008; Habeck, 2012; Tan et al., 2012)—both may be seen to recover the same algorithmic form in practice. To illustrate this equivalence, and to explain further the recursive pathway to marginal likelihood estimation, we describe each in detail below.

1.2 Reverse Logistic Regression

In the reweighting mixtures problem (cf. Geyer & Thompson 1992 and Geyer 1994) the aim is to discover an efficient proposal density for use in the importance sampling of an arbitrary target about which little is known *a priori*. Geyer's

solution was to suggest sampling not from a single density of standard form, but rather from an *ensemble* of different densities, $q_{t(j)}(\theta) = f_{t(j)}(\theta)/Z_{t(j)}$ for $j = 1, \ldots, m$ for completely known $f_{t(j)}(\theta)$ and typically unknown $Z_{t(j)}$—as may be achieved, for instance, via Metropolis-coupled MCMC (MC^3; e.g. Geyer 1992). The pooled draws, $\{\theta_i^{(j)} : i = 1, \ldots, n(j);\ j = 1, \ldots, m\}$, are then to be treated *as if* from a single pseudo-mixture density, with each free normalizing constant, and hence the appropriate weighting scheme, to be derived (up to a constant of proportionality) via RLR. For our purposes of marginal likelihood estimation, we will suppose herein that the normalization is known absolutely for at least one auxiliary/reference density in the sequence (assigned $t(1) = 0$ without loss of generality), such that $q_0(\theta) = h(\theta)$ $[Z_0 = 1]$. If $t(m) = 1$ indexes the Bayesian posterior, i.e., $q_1(\theta) = \pi(\theta)L(y|\theta)/Z$ $[Z_1 = Z]$, then the desired marginal likelihood is rendered directly from the RLR solution for this maximal temperature (Equation 11, as described below), otherwise its recovery will require one further (but trivial) importance sampling step (Equation 16). Following Geyer (1994), we write our pseudo-mixture density, $p(\theta)$, in the form,

$$p(\theta) = \sum_{s=1}^{m} [n(s)/n][f_{t(s)}(\theta)/Z_{t(s)}], \tag{7}$$

where $n(s)$ represents the sample size at each temperature and $n = \sum_{s=1}^{m} n(s)$.

Before continuing it is important to acknowledge one particular caveat on the construction of $p(\theta)$. Namely, that if labelling were available, the observed proportions in a multinomial draw from this ensemble would be unlikely to match their design constants, $n(s)/n$, though we proceed as if they would; hence, our stress of the qualifier, *pseudo*-mixture density.

With this simplification the [quasi-]log-likelihood of every $\theta_i^{(j)}$ being drawn from its true $q_{t(j)}(\theta)$ becomes

$$\log L(\{\theta_i^{(j)} : i = 1, \ldots, n(j);\ j = 1, \ldots, m\} | \{Z_{t(1)}, \ldots, Z_{t(m)}\}) = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} \log \left(f_{t(j)}(\theta_i^{(j)})/Z_{t(j)} \ /p(\theta_i^{(j)}) \right). \tag{8}$$

Setting the partial derivative in each unknown $Z_{t(k)}$ $(k = 2, \ldots, m)$ to zero yields the series of convex equations defining the RLR estimator as follows:

$$\partial/\partial Z_{t(k)} \log L(\{\theta_i^{(j)} : i = 1, \ldots, n(j);\ j = 1, \ldots, m\} | \{Z_{t(1)}, \ldots, Z_{t(m)}\}) = -\sum_{i=1}^{n(k)} 1/Z_{t(k)} - \sum_{j=1}^{m} \sum_{i=1}^{n(j)} \left([-n(k) f_{t(k)}(\theta_i^{(j)})/Z_{t(k)}^2]/[\sum_{s=1}^{m} n(s) f_{t(s)}(\theta_i^{(j)})/Z_{t(s)}] \right) \tag{9}$$

$$\to \hat{Z}_{t(k)} = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} \left(f_{t(k)}(\theta_i^{(j)})/[\sum_{s=1}^{m} n(s) f_{t(s)}(\theta_i^{(j)})/\hat{Z}_{t(s)}] \right) \tag{10}$$

$$\to \hat{Z}_{t(k)} = \sum_{i=1}^{n} \left(f_{t(k)}(\theta_i)/[\sum_{s=1}^{m} n(s) f_{t(s)}(\theta_i)/\hat{Z}_{t(s)}] \right). \tag{11}$$

The point of explicitly writing out the last step here is to highlight that having derived the 'labelled' estimator given by Equation 10 we can now 'lose the labels', (j), on our $\theta_i^{(j)}$ draws without loss of accuracy. As Kong et al. (2003) explains, "under the model as specified ... the association of draws with distribution labels is uninformative. The reason for this is that all the information in the labels for estimating the ratios is contained in the design constants, $\{n(1), \ldots, n(m)\}$".

To solve numerically for the full set of unknown $\hat{Z}_{t(k)}$ $(k = 2, \ldots, m)$ one may simply proceed via the Gauss-Seidel type recursive pathway noted earlier; that is, start with a guess of each $\hat{Z}_{t(k)}$ as input to the convex system for updating these guesses (defined by Equation 11), return the new output as input to the same, and iterate until convergence. Geyer (1994) credits Alun Thomas for suggesting this approach in the RLR context, while Kong et al. (2003) points to Vardi (1985) for a more general theoretical justification. Key asymptotic convergence proofs with necessary conditions for the uniqueness of the RLR solution are given by Geyer (1994), Chen & Shao (1997), and Kong et al. (2003). Perhaps the most important of the latter to stress is that for *identifiability*. Defining a separable sample as one for which there exist disjoint subsets P and Q of $\{1, \ldots, m\}$ such that for each θ_i and each $p \in P$ and $q \in Q$ either $f_{t(p)}(\theta_i) = 0$ or $f_{t(q)}(\theta_i) = 0$, Geyer (1994) demonstrates that the solutions to Equation 11 are unique if and only if the collection of pooled draws, $\{\theta_i : i = 1, \ldots, n\}$, is strictly *in*separable. (In fact, Geyer states uniqueness up to an additive constant, but we may drop this qualifier for the present case where $Z_0 = 1$ has been assumed.) The design of the bridging sequence so as to ensure, or at least render highly likely, the inseparability of the drawn θ_i is in effect the only constraint on our choice of ensemble—in particular, the adopted $q_{t(k)}(\theta)$ need not match the support of $\pi(\theta|y)$ individually, though their union must (Kong et al., 2003).

The gradient and Hessian of the RLR [quasi-]log-likelihood are also available analytically as

$$\partial \log L / \partial \log Z_{t(k)} = \sum_{i=1}^{n} (l\{\theta \in q_{t(k)}(\theta)\} - 1), \text{ and} \tag{12}$$

$$\partial^2 \log L / \partial \log {Z_{t(k)}}^2 = \sum_{i=1}^{n} l\{\theta \in q_{t(k)}(\theta)\}(l\{\theta \in q_{t(k)}(\theta)\} - 1), \tag{13}$$

$$\partial^2 \log L / \partial \log Z_{t(k)} \partial \log Z_{t(s)} = \sum_{i=1}^{n} l\{\theta \in q_{t(k)}(\theta)\} l\{\theta \in q_{t(s)}(\theta)\}, \tag{14}$$

respectively, where $l\{\theta \in q_{t(k)}(\theta)\} = f_{t(k)}(\theta)[n(s)/n]/Z_{t(k)}/p(\theta)$. Aside from offering a non-recursive pathway to Z via downhill gradient search from a starting position sufficiently near to the true $Z_{t(k)}$ (as implemented by Tan et al. 2012, for example, using a trust region algorithm in `R`), the above equations also facilitate an estimate of the *uncertainty* associated with the recovered $\log \hat{Z}_{t(k)}$. Geyer (1994) gives a generic limiting form for this uncertainty in his asymptotic Normality theorem for the RLR, while Kong et al. (2003) and Tan et al. (2012) offer more practical matrix constructions (the latter designed to avoid the burden of generalized matrix inversion).

A 'naïve' *alternative derivation* of the RLR estimator relevant to the thermodynamic integration via importance sampling methodology we describe in Section

2 is that given by Jiang & Tanner (2003) in their discussion of Kong et al. (2003): simply take $p(\theta)$ as a pseudo-importance sampling density for each $q_{t(k)}(\theta)$, such that

$$Z_{t(k)} = \int_{\Omega(\theta)} f_{t(k)}(\theta)/p(\theta)\{p(\theta)d\theta\}, \tag{15}$$

and solve recursively,

$$\hat{Z}_{t(k)} = \sum_{i=1}^{n} \left(f_{t(k)}(\theta_i)/p(\theta_i)/n \right) \text{ for } \theta_i \sim p(\theta). \tag{16}$$

The finite variance of the RLR estimator under this construction may be seen as a consequence of its implicit use of defensive importance sampling (Hesterberg, 1991) whereby every target density is itself contained within the proposal density. The (original) stabilized harmonic mean estimator (SHME; $\hat{p}_3$) of Newton & Raftery (1994) is in fact based on exactly the approach described by Jiang & Tanner, but with a further step of normalization by the sum of weights (serving only to speed up its iterative convergence). If $\pi(\theta|y)$ is not expressly contained within the proposed $q_{t(k)}(\theta)$ family, then as mentioned earlier a final importance sampling step will be needed to recover the desired Z—for this the relevant formula is given by Equation 16 with $f_{t(k)}(\theta_i)$ replaced by $\pi(\theta_i)L(y|\theta_i)$.

1.3 The Density of States

Yet another construction of the convex series of $Z_{t(k)}$ updates characterizing the RLR appoach (cf. Equation 11) has recently been demonstrated in the context of free energy estimation for molecular interactions by Habeck (2012) and Tan et al. (2012). In this alternative method rather than aiming directly for estimation of the marginal likelihood one aims instead to reconstruct a closely-related measure, the density of states (DOS), $g(e)$, "defined" in terms of a composition of the Dirac delta 'function', $\delta(\cdot)$, as

$$g(e) = \int \pi(\theta)\delta(e + \log L(y|\theta))d\theta. \tag{17}$$

Our somewhat uncharitable placement of quotation marks in the above is intended to draw attention to a crucial point overlooked in these previous studies; namely, that the composition of the Dirac delta 'function'—which is itself not strictly a function, being definable only as a *measure* or a *generalized function*—lacks an intrinsic definition. Hörmander (1983) proposes a definition in $\mathbb{R}^n$ valid only when the composing function, here $v(\theta) = E + \log L(y|\theta)$, is continuously differentiable and $dv(\theta)/d\theta$ nowhere zero; the latter condition in particular must be considered problematic in the general case that $L(y|\theta)$ features a global maximum on $\Omega(\theta)$, and disasterous in the specific case that $L(y|\theta)$ features a constant set of non-zero measure with respect to $\pi(\theta)$! Here we suggest an alternative definition of the DOS as the derivative of an inverse survival function, which elucidates its connection to nested sampling.

Chopin & Robert (2010) characterize *nested sampling* by supposing first the invertibility of the survival function, $\phi^{-1}(l)$, where

$$\phi^{-1}(l) = \text{pr}\{L(y|\theta) > l\} \text{ for } \theta \sim \pi(\theta), \tag{18}$$

which they ensure by stipulating $L(y|\theta)$ continuous and the support of $\pi(\theta)$ connected. With $\phi(\theta)$ and $\phi^{-1}(l)$ so defined the nested sampling identity may be written as

$$Z = \int_0^1 \phi(x)dx. \tag{19}$$

Now, consider the distribution function defined in terms of the free energy, $E(\theta) = -\log L(y|\theta)$ (i.e., $e = -\log l$), such that

$$G(e) = \text{pr}\{E(\theta) < e\} \text{ for } \theta \sim \pi(\theta), \tag{20}$$

from which the DOS may be defined for differentiable $G(e)$ (again, $L(y|\theta)$ can have no set of constant values of non-zero measure) as

$$g(e) = dG(e)/de. \tag{21}$$

Since $G(e) = \phi^{-1}(\exp[-e])$ we have

$$g(e) = d\phi^{-1}(\exp[-e])/de = d\phi^{-1}(l)/dl \times dl/de = d\phi^{-1}(l)/dl \times -\exp[-e]. \tag{22}$$

Transforming the nested sampling identity of Equation 19 by substitution with $\phi^{-1}(l)$ yields

$$Z = \int_{\phi^{-1}(\infty)}^{\phi^{-1}(0)} \phi(x)dx = \int_{\infty}^{0} \phi(\phi^{-1}(l)) \times d\phi^{-1}(l)/dl \times dl, \tag{23}$$

and hence,

$$Z = \int_{\infty}^{0} l \times -g(e)\exp[e] \times dl = \int_{-\infty}^{\infty} \exp[-e] \times -g(e)\exp[e] \times dl/de \times de. \tag{24}$$

That is, consistent with the requirements of Habeck (2012) and Tan et al. (2012), our DOS formulation returns the identity,

$$Z = \int_{-\infty}^{\infty} g(e)\exp[-e]de. \tag{25}$$

To make use of this identity Habeck (2012) suggests sampling from a series of bridging densities indexed by $t(k)$ as in the RLR method but with each featuring an explicit dependence on the free energy, $-\log L(y|\theta)$, such that $q_{t(k)}(\theta) = f^*_{t(k)}(E(\theta))\pi(\theta)/Z_{t(k)}$ [here we write $f^*(E(\theta))$ to distinguish such specifically energy-dependent bridging sequences from their more general $f(\theta)$ counterparts in "ordinary" RLR]. Accordingly each $Z_{t(k)}$ may be written in terms of the DOS as

$$Z_{t(k)} = \int_{\Omega(\theta)} f^*_{t(k)}(E(\theta))\pi(\theta)d\theta = \int_{-\infty}^{\infty} g(e)f^*_{t(k)}(e)de. \tag{26}$$

Treating the pool of energies, $\{E_i = E(\theta_i^{(j)}) : i = 1, \ldots, n(j); j = 1, \ldots, m\}$, corresponding to the pool of θ_i draws, as a single simulation of length, n, from the

pseudo-mixture, $p(e) = \sum_{s=1}^{m}[n(s)/n][g(e)f^*_{t(s)}(e)/Z_{t(s)}]$, one may construct directly the recursive importance sampling estimator of each normalizing constant,

$$\begin{aligned}\hat{Z}_{t(k)} &= \sum_{i=1}^{n} g(E_i)f^*_{t(k)}(E_i)/p(E_i)/n \\ &= \sum_{i=1}^{n}\left(g(E_i)f^*_{t(k)}(E_i)/[\sum_{s=1}^{m}[n(s)g(E_i)f^*_{t(s)}(E_i)/Z_{t(s)}]\right) \\ &= \sum_{i=1}^{n}\left(f^*_{t(k)}(E_i)/[\sum_{s=1}^{m}[n(s)f^*_{t(s)}(E_i)/Z_{t(s)}]\right).\end{aligned} \tag{27}$$

Surprisingly, despite having based our derivation of $\hat{Z}_{t(k)}$ on the DOS we never need know or even directly estimate $g(e)$ as it cancels out in the final step! Inspection of Equation 27 confirms that the DOS solution for $\hat{Z}_{t(k)}$ $(k = 2, \ldots, m)$ matches exactly that of the RLR with $f_{t(k)}(\theta) = f^*_{t(k)}(E(\theta))\pi(\theta)$.

As Habeck (2012) has insightfully highlighted, despite the restriction that the bridging distributions of the DOS take the form $q_{t(k)}(\theta) = f^*_{t(k)}(E(\theta))\pi(\theta)/Z_{t(k)}$, the sampling schemes of both thermodynamic integration along the power posteriors path and nested sampling can be easily accommodated within this framework given astute choices for $f^*_{t(k)}(E(\theta))$—namely, $f^*_{t(k)}(E(\theta)) = \exp[-t(k)E(\theta)]$ for the former and $f^*_{t(k)}(E(\theta)) = H(E_{t(k)} - E(\theta))$ with $H(\cdot)$ the Heaviside step function and $E_{t(k)} = E(\theta_1^{(k-1)})$ for the latter. In principle one could even pool draws from both schemes together under the DOS framework to construct a joint power posteriors–nested sampling approximation to the marginal likelihood—with the sub-sample estimates of Z based on each separately providing a check on the computations leading to their combined solution; though the computational complexity of such a scheme is unlikely to be attractive to many users. Finally, it is perhaps also important to note here that our definition of $G(e)$ as a probability with respect to $\pi(\theta)$ may of course be relaxed to become a probability with respect to some generic reference distribution, $h(\theta)$, [assumed closer to $\pi(\theta|y)$ than $\pi(\theta)$] as a means of improving the efficiency of such DOS-based integration.

1.4 Prior-Sensitivity Analysis

In the Bayesian framework (Jeffreys, 1961; Jaynes, 2003) the ratio of marginal likelihoods under rival hypotheses (i.e., the Bayes factor) operates directly on the prior odds ratio for model selection, as

$$\begin{aligned}\pi(M_1|y)/\pi(M_2|y) &= [\pi(M_1)/\pi(M_2)][\pi(y|M_1)/\pi(y|M_2)] \\ &= Z_{M(1)}/Z_{M(2)}\pi(M_1)/\pi(M_2).\end{aligned} \tag{28}$$

A much maligned feature of the marginal likelihood in this context is its possible sensitivity to the choice of the parameter priors, $\pi(\theta|M_1)$ and $\pi(\theta|M_2)$. When there is limited information or theoretical motivation available to inform this choice the resulting Bayes factor can appear arbitrary in value. (On the other hand, viewed as a quantitative implementation of Ockham's Razor, the key role of prior precision may well serve as strong justification for the use of Bayesian model selection in the scientific context; cf. Jeffreys & Berger 1991.) In their

influential treatise on this topic Kass & Raftery (1995) thus argue that some form of prior-sensitivity analysis be conducted as a routine part of all Bayesian model choice experiments; their default recommendation being the recomputation of the Bayes factor under a doubling and halving of key hyperparameters.

If the initial marginal likelihoods have been computed under an amenable sampling scheme then, as Chopin & Robert (2010) point out for the case of nested importance sampling, alternative Bayes factors under (moderate) prior rescalings may be easily recovered by appropriately reweighting the existing draws, without the need to incur further (computationally expensive) likelihood function calls; and indeed the RLR method was developed specifically to facilitate such computations (though in the reweighting mixtures context; Geyer & Thompson 1992; Geyer 1994). That is, having recovered estimates for each $\hat{Z}_{t(k)}$ under our nominal prior the pooled draw pseudo-mixture density, $p(\theta) = \sum_{k=1}^{m}[n(k)/n]q_{t(k)}(\theta)$, now offers by design an efficient proposal for importance sampling of many other targets near the posterior. The alternative marginal likelihood estimate, $\hat{Z}_{\text{alt}}$, under alternative prior, $\pi_{\text{alt}}(\theta)$, simply becomes

$$\hat{Z}_{\text{alt}} = \sum_{i=1}^{n} L(y|\theta_i)\pi_{\text{alt}}(\theta_i)/p(\theta_i)/n. \tag{29}$$

The stability of this importance sampling procedure may then be monitored via the effective sample size, $\text{ESS} = n/[1 + \text{var}_p\{\pi_{\text{alt}}(\theta)/p(\theta)\}]$, following Kong et al. (1994).

2. THERMODYNAMIC INTEGRATION VIA IMPORTANCE SAMPLING

Inspired by the recursive pathway (of the RLR and DOS) we present here yet another such strategy for marginal likelihood estimation, which we name, 'thermodynamic integration via importance sampling' (TIVIS). Although quite novel at face value it is easily shown to be yet another manifestation of the RLR methodology; yet by effectively recasting the RLR as a thermodynamic integration procedure we attain insight into the relationship between its error budget and the choice of bridging sequence. Specifically, the error in the estimation of each $Z_{t(k)}$ may be thought of as dependent on both the J-divergence (Lefebvre et al., 2010) between it and the remainder of the ensemble (via the thermodynamic identity) and on the accuracy of our estimates for those other $Z_{t(j)}$ $(j \neq k)$. Thus, we suggest that an effective choice of bridging sequence will produce a near equal spacing of $\log \hat{Z}_{t(k)}$ (a proposal we explore numerically in Section 3 below).

To construct the TIVIS estimator we once again assume the availability of pooled draws, $\{\theta_i^{(j)} : i = 1, \ldots, n(j);\ j = 1, \ldots, m\}$, from a sequence of bridging densities, $q_{t(j)}(\theta) = f_{t(j)}(\theta)/Z_{t(j)}$ $(j = 1, \ldots, m)$, with each $f_{t(k)}(\theta)$ known exactly. Moreover, we suppose that $t(1) = 0$ indexes a normalized reference/auxiliary, $\pi(\theta)$ or $h(\theta)$, such that $Z_1 = 1$ is known, but with the remaining $Z_{t(k)}$ typically unknown. Despite our subsequent use of the thermodynamic identity, however, we do not require here that the bridging densities follow the geometric path between these two extremes. Now, rather than seek each $\hat{Z}_{t(k)}$ via direct importance sampling from $p(\theta)$ as per the RLR, the TIVIS method is to instead seek each normalization constant via thermodynamic integration from its preceding density

in the ensemble, $q_{t(k-1)}(\theta)$, using the identity,

$$\log Z_{t(k)} = \int_0^1 E_{\pi_x^k}\{\log\left(f_{t(k)}(\theta)/q_{t(k-1)}(\theta)\right)\}dx, \tag{30}$$

where $\pi_x^k(\theta) \propto [q_{t(k)}(\theta)]^x[q_{t(k-1)}(\theta)]^{1-x} \propto [f_{t(k)}(\theta)]^x[f_{t(k-1)}(\theta)]^{1-x}$. (Here we have adopted the unconventional notation, x, for the thermodynamic integration variable to avoid confusion with the $t(k)$ of the bridging densities.) For existence of the log-ratio in Equation 30 we must impose the strict condition (*not* necessary for ordinary RLR) that all $q_{t(k)}(\theta)$ share matching supports. Importance sampling from $p(\theta)$ allows construction of the appropriate (but unnormalized) weighting function,

$$w(\theta, x) = [f_{t(k)}(\theta)]^x[f_{t(k-1)}(\theta)]^{1-x}/p(\theta), \tag{31}$$

which in substitution to Equation 30 yields the TIVIS estimator,

$$\log(\hat{Z}_{t(k)}/\hat{Z}_{t(k-1)}) = \int_0^1 [\sum_{i=1}^n \log\left(f_{t(k)}(\theta_i)/f_{t(k-1)}(\theta_i)\right) w(\theta_i, x)]/[\sum_{i=1}^n w(\theta_i, x)]dx. \tag{32}$$

In computational terms, numerical solution of the one-dimension integral in the above may be achieved to arbitrary accuracy by simply evaluating the integrand at sufficiently many x_j on the unit interval, followed by summation with Simpson's rule. If the sequence of bridging densities is well-chosen (and suitably ordered) the J-divergence between each $q_{t(k)}(\theta)$ and $q_{t(k-1)}(\theta)$ pairing should be *far less* than that between prior and posterior, such that a naïve regular spacing of the x_j will suffice.

To show the equivalence between this estimator and that of the RLR defined by Equation 11 we simply observe that the derivative of the denominator in Equation 32 equals the numerator, and thus by analogy to $\int_0^1 u'(x)/u(x)dx = \log u(1) - \log u(0)$ we have

$$\log(\hat{Z}_{t(k)}/\hat{Z}_{t(k-1)}) = \log[\sum_{i=1}^n f_{t(k)}(\theta_i)/p(\theta_i)] - \log[\sum_{i=1}^n f_{t(k-1)}(\theta_i)/p(\theta_i)]$$

$$\rightarrow \log(\hat{Z}_{t(k)}/\hat{Z}_{t(k-1)}) = \log[\sum_{i=1}^n f_{t(k)}(\theta_i)/p(\theta_i)/n] - \log[\sum_{i=1}^n f_{t(k-1)}(\theta_i)/p(\theta_i)/n]$$

$$\rightarrow \log \hat{Z}_{t(k)} = \log[\sum_{i=1}^n f_{t(k)}(\theta_i)/p(\theta_i)/n]. \tag{33}$$

In the following two case studies we explore by numerical example both the diversity of computational implementions of the recursive pathway (with particular reference to their efficiency in $L(y|\theta)$ calls; Section 3) and the utility of these RLR-normalized bridging sequences for testing prior-sensitivity (Section 4).

3. CASE STUDY: BANANA-SHAPED PSEUDO-LIKELIHOOD FUNCTION

For our first case study we demonstrate application of the recursive pathway to estimation of the marginal likelihood for a banana-shaped pseudo-likelihood

function [hence, we write here $L(\theta)$ instead of $L(y|\theta)$] in two-dimensions, defined as

$$L(\theta = \{\theta_1, \theta_2\}') = \exp[-(10 \times (0.45 - \theta_1))^2/4 - (20 \times (\theta_2/2 - \theta_1^4))^2)], \tag{34}$$

with a Uniform prior density of $\pi(\theta) = 1/4$ on the rectangular domain, $[-0.5, 1.5] \times [-0.5, 1.5]$. A simple illustration of this $L(\theta)$ as a (logarithmically-spaced) contour plot is presented in the left-hand panel of Figure 1. Brute-force numerical integration via quadrature returns the "exact" solution, $Z = 0.01569[6]$ (or $\log Z = -4.154[3]$).

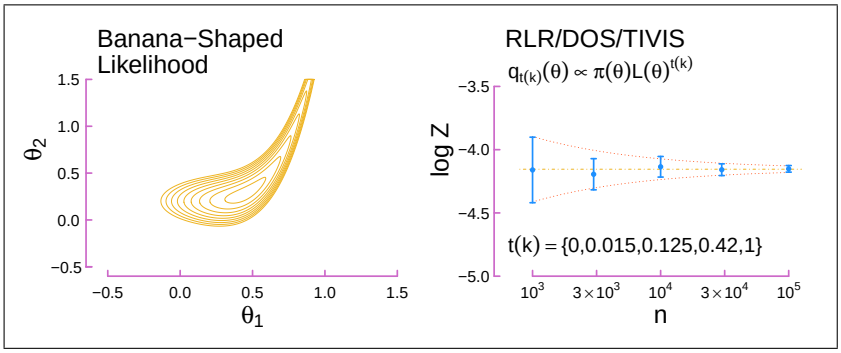

FIG 1. *The banana-shaped pseudo-likelihood function of our first case study (Equation 34 of Section 3) is illustrated graphically here (left-hand panel) as a (logarithmically-spaced) contour plot on the domain of our Uniform prior,* $[-0.5, 1.5] \times [-0.5, 1.5]$. *Convergence of the RLR/DOS/TIVIS estimator for the corresponding marginal likelihood under Metropolis-coupled MCMC sampling of the power posterior (at five pre-specified temperatures) as a function of the total sample size is shown in the right-hand panel. The marked points and error bars on this Figure indicate respectively the recovered mean and standard error in* $\log \hat{Z}$ *for thirty trials at each* n. *The dash-dotted, yellow line indicates the "exact"* $\log Z$ *for this example derived via brute-force quadrature, and the dotted, red lines highlight the* $\sim 1/\sqrt{n}$ *convergence rate.*

As a benchmark of the method we first apply the RLR/DOS/TIVIS estimator to samples drawn from a sequence of bridging densities following the power posteriors path. Though even a cursory inspection of the pseudo-likelihood function for this simple case study is sufficient to confirm its unimodality and to motivate a family of suitable proposal densities for straight-forward importance sampling of $\pi_t(\theta) \propto \pi(\theta)L(\theta)^t$, for illustrative purposes we have chosen to implement an MC^3 (Geyer, 1994) approach here instead; the latter being ultimately amenable to a much wider variety of Bayesian analysis problems than the former. With regard to the choice of tempering schedule, for which we first trial a thermodynamic integration-inspired scaling, we refer to Gelman & Meng (1998), who note pragmatically that *a priori* estimation of the optimal $\pi(t)$ will often be far more difficult than the estimation of Z itself; thus we simply follow Friel & Pettitt (2008) in adopting a pre-determined set of (five) temperatures spaced geometrically as $t = \{(0, 1/4, 1/2, 3/4, 1)^3\}$.

To quantify the behavior of the RLR/DOS/TIVS estimator under this power posteriors sampling strategy we have run the above computational experiment thirty times at each of five total sample sizes spaced as $n = \{10^3, 3 \times 10^3, 10^4, 3 \times 10^4, 10^5\}$; the results of which are illustrated in the right-hand panel of Figure 1. Code for reproducing our experiment using the RLR/DOS and TIVIS schemes alternately (cf. Sections 1.2/1.3 and 2) is available from the first author upon

request. As demonstrated by Geyer (1994), whether for Metropolis-coupled or regular MCMC sampling, the RLR estimator will converge asymptotically to the true $\log Z$ with a standard error decreasing as $1/\sqrt{n}$; and indeed we can confirm that such a scaling with n is already evident in the standard errors recovered from the present example (see Figure 1). More importantly, we can also confirm a close agreement between this mean standard error (computed from repeat simulation) and the estimated asymptotic standard deviation of $\log \hat{Z}$ computed using the matrix form from Kong et al. (2003) [their Equation 4.2], highlighting the utility of the latter for efficient uncertainty estimation.

Finally, we can use the derived sequence of $\log \hat{Z}_{t(k)}$ to check the suitability of our adopted tempering schedule for the present case study; with the expectation according to our TIVIS formulation that an efficient pathway will produce a near equal spacing. In this example for $t = \{(0, 1/4, 1/2, 3/4, 1)^3\}$ we recover $\log \hat{Z}_{t(k)} \approx \{0, -0.9, -2.3, -3.3, -4.2\}$—i.e., a spacing of $\{-0.9, -1.4, -1, -0.9\}$ with sample standard deviation, $\hat{\sigma}_{\log \hat{Z}_k} = 0.24$—giving a standard deviation of $\log \hat{Z}$ about the true $\log Z$ of $\hat{\sigma}_{\hat{Z}} = 0.10$ at $n = 10,000$. By way of comparison, for alternative temperature exponents, $\{1, 2, 4, 5\}$, we recover $\hat{\sigma}_{\log \hat{Z}_k} = \{1.1, 0.44, 0.38, 0.55\}$ with $\hat{\sigma}_{\hat{Z}} = \{0.19, 0.16, 0.12, 0.13\}$—confirming that our thermodynamic integration-inspired, $\hat{\sigma}_{\log \hat{Z}_k}$-minimizing choice of temperature sequence was indeed most efficient.

With this power posteriors version of RLR as benchmark we now consider the merits of two alternative schemes for defining, and sampling from, the required sequence of bridging densities, $q_{t(k)}(\theta)$, in Sections 3.1 and 3.2 below.

3.1 Thermodynamic Integration from a Reference/Auxiliary Density

As highlighted by Lefebvre et al. (2010), the error budget of thermodynamic integration over the geometric path depends to first-order upon the J-divergence between the reference/auxiliary density, $h(\theta)$, and the target, $\pi(\theta|y)$. Thus, it will generally be more efficient to set a 'data-driven' $h(\theta)$—such as may be recovered from the position and local curvature of the posterior mode—than to integrate 'naïvely' from the prior, i.e., $h(\theta) = \pi(\theta)$. Here we demonstrate the corresponding improvement to the performance of the RLR/DOS/TIVIS estimator resulting from the relevant choices, $h(\theta) \sim \mathcal{T}_{\text{Trunc.}}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}})$ and $h(\theta) \sim \mathcal{N}_{\text{Trunc.}}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}})$. Here $\mathcal{T}_{\text{Trunc.}}$ and $\mathcal{N}_{\text{Trunc.}}$ denote the two-dimensional Student's t ($\nu = 1$) and Normal distributions (truncated to our prior support), respectively, while μ_{mode} denotes the posterior mode and Σ_{mode} its local curvature (recovered here analytically, but estimable at minimal cost in many Bayesian analysis problems via standard numerical methods). As before we apply MC^3 to explore the tempered posterior and run both experiments thirty times at each of our five n. In contrast to the power posteriors case we adopt here a regular temperature grid, $t = \{0, 0.25, 0.5, 0.75, 1\}$, to allow for the imposed/intended similiarity between $\pi(\theta|y)$ and $h(\theta)$. Our results are presented in Figure 2 and discussed below.

As expected from both theoretical considerations (Gelman & Meng, 1998; Lefebvre et al., 2010), and reports of practical experience with other marginal likelihood estimators (Fan et al., 2012), use of a 'data-driven' auxiliary in this example has indeed reduced markedly the standard error of the RLR/DOS/TIVIS scheme (at fixed n) with respect to that of the naïve (power posteriors) path,

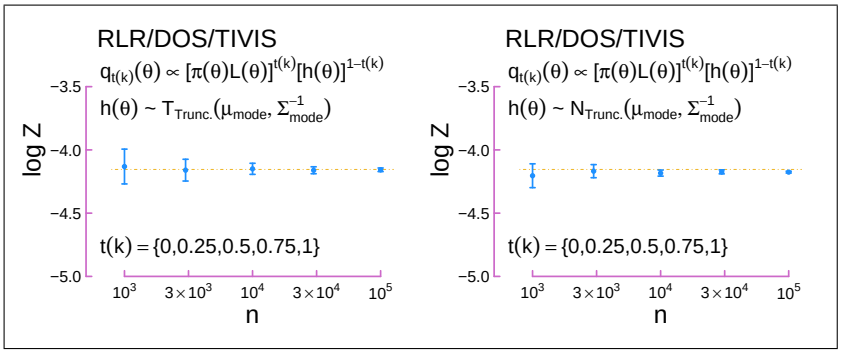

FIG 2. *Convergence of the RLR/DOS/TIVIS estimator for the marginal likelihood of our banana-shaped pseudo-likelihood function under Metropolis-coupled MCMC sampling (at five pre-specified temperatures) on the geometric path between a 'data-driven' reference/auxiliary density, $h(\theta)$, and the posterior, shown as a function of the total sample size. The adopted $h(\theta)$ takes a two-dimensional Student's t form in the left panel and a Normal form in the right, with its controlling parameters (μ_{mode} and Σ_{mode}) in each case set to the location and curvature of the posterior mode. A marked reduction in standard error (at fixed n) with respect to that of the naïve (power posteriors) path, i.e., $h(\theta) = \pi(\theta)$, is evident from comparison with Figure 1.*

i.e., $h(\theta) = \pi(\theta)$. Moreover, in this instance the (thinner-tailed) Normal auxiliary form has out-performed the (fatter-tailed) Student's t (one d.o.f.), which is again consistent with theoretical expectations as a quick computation using the "exact" $\log Z$ confirms $J[\mathcal{N}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}}), h(\theta)] \ll J[\mathcal{T}(\mu_{\text{mode}}, \Sigma^{-1}_{\text{mode}}), h(\theta)]$. (Unfortunately, no such 'precise' optimisation of this form for $h(\theta)$ is possible *a priori* without knowledge of the desired Z; and even a crude estimator of the J-divergence run with, e.g., the Laplace approximation to the marginal likelihood will nevertheless add numerous extra likelihood evaluations to the computational budget.)

3.2 Ellipse/Ellipsoid-Based Nested Sampling

Recalling our derivation of the DOS pathway to the marginal likelihood, which began with the fundamental identity of the rival nested sampling algorithm (see Section 1.3 and Equation 19 above), it is of some interest to compare directly the performance of these two particular methodologies. The present case study with its Uniform prior density is in fact well suited to this purpose since in the field of cosmological model selection, where nested sampling has been most extensively-used of late (Mukherjee et al., 2006; Feroz & Hobson, 2008), it is standard practice to adopt separable priors from which a Uniform sample space may be easily constructed under the simple quantile function transformation; which, for the discussion below, we assume has been done such that $\pi(\theta)$ may be taken as strictly Uniform on $[0, 1]^N$. Given these conditions Mukherjee et al. (2006) outline a crude-but-effective scheme for exploring the constrained-likelihood space of nested sampling, in which the new "live" particle for each update must be drawn with density proportional to $\pi(\theta)I(L(\theta) > L(\theta_{i-1}))$.

Under the Mukherjee et al. (2006) scheme, to draw the required θ_i one simply: (I) identifies the minimum bounding ellipse (or with $D(\theta) > 2$, the minimum bounding *ellipsoid*) for the present set of "live" particles; (II) expands this ellipse by a small factor ∼1.5-2 with the aim of enclosing the full support of $I(L(\theta) > L(\theta_{i-1}))$; and (III) draws randomly from its interior until a valid

$\{\theta_i, L(\theta_i)\}$ is discovered. Supposing the elliptical sampling window thus defined has been enlarged sufficiently to fully enclose the desired likelihood surface (which it must do to ensure unbiased sampling of $\{\theta_i, L(\theta_i)\}$, although we can rarely be *sure* that it has) it remains unlikely to match its shape exactly, leading to an overhead of n_{oh} discarded draws, $\{\theta_i^{(j)} : L(\theta_i^{(j)}) < L(\theta_{i-1}),\ j = 1, \ldots, n_{\text{oh}}\}$. At each θ_i the incurred n_{oh} may be thought of as a single realization of the negative binomial distribution with p equal to the fraction of the bounded ellipse for which $L(\theta) < L(\theta_{i-1})$; hence, $E(n_{\text{oh}}) = 1/p - 1$. The magnitude of this overhead can in general be expected to scale with the geometric volume of the parameter space, potentially limiting the utility of this otherwise dimensionally-insensitive Monte Carlo-based estimator. However, where applicable the Mukherjee et al. (2006) scheme may well prove more efficient than the alternative of constrained-MCMC-sampling (cf. Friel & Wyse 2012) in which one must discard at least ∼10-20 burn-in moves (each with a necessary $L(\theta)$ call) per accepted θ_i to achieve approximate stationarity.

Applying this ellipse-based approach to nested sampling of the pseudo-likelihood function of Equation 34 (with $n/7$ live particles in each case and a small extrapolation of the mean L_{live} times $\exp(-7)$ at the final step; cf. Skilling 2006) we recover a convergence to the true $\log Z$ (as shown in Figure 3) of efficiency (in n) comparable to that of the RLR/DOS/TIVIS estimator run over the geometric path with Student's t auxiliary (Figure 2, left-hand panel). However, the overhead of $n_{\text{oh}} \sim 3.4$ total likelihood calls observed here on average per accepted θ_i should be a concern for applications of nested sampling in which the likelihood function may be genuinely expensive to evaluate; indeed for modern cosmological simulations MCMC exploration of the $D(\theta) \lesssim 12$ posterior is effectively a super-computer-only exercise due solely to the cost of solving for $L(y|\theta)$. [At this point the skeptical reader might suggest that the distinctly non-elliptical $L(\theta)$ considered in this example be considered a particularly "unfair" case for testing the Mukherjee et al. (2006) method, but such banana-shaped likelihoods are in fact quite common in higher-order cosmological models; see, for instance, Davis et al. 2007.] We therefore suggest that in general one might improve upon the efficiency of ellipse-based nested sampling by co-opting its bridging sequence into the RLR framework under 'soft' sampling of the likelihood constraint, and in the following we give one such specific example.

By 'soft' sampling we simply mean sampling not subject to hard likelihood thresholding; the simplest version of which would be to follow the original nested sampling path, drawing as usual from each proposal ellipse until a suitable replacement point is discovered *but without discarding the* n_{oh} *lower likelihood draws at each step*. The resulting $q_{t(k)}(\theta)$ ensemble for $k > 1$ then takes the form, $q_{t(k)}(\theta) \propto I(\theta \in Ell[E_{\text{live}(k)}])$, instead of $q_{t(k)}(\theta) \propto I(\theta \in Ell[E_{\text{live}(k)}]) \times I(L(\theta_i) > L(\theta_{i-1}))$, where $Ell[E_{\text{live}(k)}]$ denotes the minimum bounding ellipse/ellipsoid for the current set of live particles. Computational experiments on the present case study for $N = 142$ live particles ($n \approx 1000$) confirm the superior accuracy of this approach over direct nested sampling; with a mean and standard error from repeat simulation of $\log \hat{Z} = -4.15 \pm 0.09$ for the former, and -4.13 ± 0.15 for the latter. However, with each nested sampling step contributing its own $q_{t(k)}(\theta)$ the computational burden of solving for the full set of $\hat{Z}_{t(k)}$ under RLR quickly becomes prohibitive at larger n.

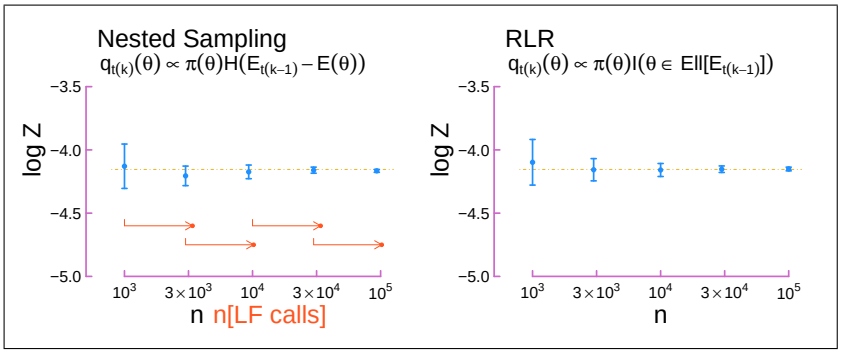

FIG 3. *The performance of nested sampling (left panel) as a marginal likelihood estimator for our banana-shaped pseudo-likelihood function, run under the ellipse-based strategy for exploring the sequence of constrained-likelihood densities proposed by Mukherjee et al. (2006); compared with that of the RLR/DOS (right panel) for our alternative scheme of 'soft' ellipse-based sampling (see Section 3.2). The two schemes converge to the true* $\log Z$ *at a similar rate in* n *(the number of draws from intermediate distributions compromising the final sample), but the former incurs a factor of* $\sim$*3.4 more likelihood function calls in this case, owing to the often imprecise match between the proposal ellipse and the true constrained likelihood surface. This overhead of ellipse-based nested sampling can be expected to grow with the geometric volume; though for problems amenable to this technique (i.e., with separable priors) will likely remain preferable to the MCMC-based alternative (cf. Friel & Wyse 2012).*

Therefore, we propose the alternative strategy of moving up in fixed quantiles of the sampled likelihoods. For instance, (I) begin with $n(1)$ draws from $q_{t(k=1)=0}(\theta) = \pi(\theta)$ [with $\pi(\theta)$ the unit hypercube under the quantile transformation], (II) take a further $n(2)$ draws via simple rejection sampling from $q_{t(k=2)}(\theta) \propto I(\theta \in Ell[E_{t(k-1)}])$ where $Ell[E_{t(k-1)}]$ now denotes the minimum bounding ellipse/ellipsoid for the upper third quantile of the previous $E_{t(k-1)}$ draws, and (III) repeat (II) a further $m-2$ times. Much like nested sampling itself, for unimodal likelihoods this scheme will clearly act to evolve a series of nested ellipses in from the prior to the vicinity of the posterior mode; with modification for the multi-modal case achievable via, e.g., k-means clustering, as per Feroz & Hobson (2008)'s implementation of the Mukherjee et al. (2006) scheme. Running this algorithm for ten steps with equal $n(k)$ we are able to recover the marginal likelihood for our banana-shaped pseudo-likelihood function with very similar efficiency in n to that of nested sampling (as shown in Figure 3)—and thus ($\sim$3 times) *superior efficiency in the number of total likelihood calls.*

Note that even if the prior were not amenable to the quantile transformation one could still apply our ellipse-based RLR method with the same efficiency in $L(\theta)$ calls simply by running MCMC, without calls to $L(\cdot)$, to explore $\pi(\theta)I(\theta \in Ell[E_{t(k-1)}])$ at each $k > 1$ instead of the rejection sampling used in the present example. Indeed, the irony is that this may be in fact the *only* truly interesting case for RLR in such ellipse-based nested schemes since the normalization of every such $q_{t(k)}(\theta)$ is otherwise trivially computable as the volume of the corresponding ellipsoid, V_s. These directly-computed $Z_{t(k)}$ [$= V_s$] can of course be employed directly to estimate Z as in the final importance sampling step for the RLR (cf. Section 1.2), treating the pooled θ_i as if from the pseudo-mixture density,

$$p(\theta) = \sum_{s=1}^{m} [n(s)/n][I(\theta \in Ell[E_{\text{live}(s)}])/V_s]. \tag{35}$$

For the $N = 142$ ($n \approx 1000$) experiment considered above this pseudo-mixture density scheme yields a surprisingly accurate mean and standard error of $\log \hat{Z} = -4.155 \pm 0.015$ (compared to -4.15 ± 0.09 for the RLR). Further computational experiments are now underway to better quantify the advantages offered by this approach to harnessing the information content of these otherwise discarded draws in the ellipse-based nested sampling paradigm (Feroz et al., in prep.).

4. CASE STUDY: NORMAL MIXTURE MODELING OF THE GALAXY DATASET

The well-known galaxy dataset, first proposed as a test case for kernel density estimation by Roeder (1990), consists of precise recession velocity measurements (in units of 1000 km s^{-1}) for 82 galaxies in the Corona Borealis region of the Northern sky reported (as "the unfilled sample") by Postman et al. (1986). The purpose of the original astronomical study was to search—in light of a then recently discovered void in the neighboring Boötes field (Kirshner et al, 1981)—for further large-scale inhomogeneities in the distribution of galaxies, as evidence for the operation of hierarchical clustering processes on cosmological scales (Gunn & Gott, 1972). Given the well-defined selection function of their survey Postman et al. (1986) were easily able to compute as benchmark the recession velocity density function expected under the null hypothesis of a uniform distribution of galaxies throughout space; and by visual comparison of this density against a histogram of their observed velocities the astronomers were able to establish strong evidence against the null, as was their aim.

However, as Roeder (1990) soon realized, under this favored hypothesis of an inhomogeneous galaxy distribution one can pose the far more challenging statistical question of "how many distinct clustering components are in fact present in the recession velocity dataset"? Many authors have since attempted to answer this question as a univariate, Normal mixture modeling problem, with notable contributions in the Bayesian framework including those of Escobar & West (1995), Phillips & Smith (1996), Richardson & Green (1997) and Stephens (2000). The pre-Millennial contributions to this end being well summarized by Aitkin (2001), who highlights the extreme sensitivity to the specified priors of the inferred number of components at the posterior mode. For this reason the galaxy dataset provides a most interesting test case for the utility of the recursive pathway with prior-sensitivity analysis as described in Section 1.4.

In the following we detail the Normal mixture model adopted (Section 4.1), discuss various astronomical motivations for our priors (Section 4.2), run RLR on this problem with MC^3 exploration of the posterior (Section 4.3), and explore the prior-sensitivity of our results with a comparison to previous analyses (Section 4.4).

4.1 Normal Mixture Model

Following Richardson & Green (1997) we write the k-component Normal mixture model with component weights, w, in the latent allocation variable form for data vector, y, and (unobserved) allocation vector, z, such that

$$\pi(z_i = j) = w_j \text{ and } \pi(y_i | z_i = j) = f_{\mathcal{N}}(y_i | \theta_j). \tag{36}$$

Here $f_{\mathcal{N}}(\cdot|\theta_j)$ represents the one-dimensional Normal density, which we will reference in mean–precision syntax as $\mathcal{N}(\mu_j, \tau_j^{-1})$, i.e., $\theta_j = \{\mu_j, \tau_j\}'$. The likelihood

function for independent y_i is thus recoverable via summation over the unobserved, z_i, as

$$L(y|\theta, w, k) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f_{\mathcal{N}}(y_i|\theta_j). \tag{37}$$

Given priors for the number of components in the mixture, the weights at a given k, and the vector of mean–precisions—i.e., $\pi(k)$, $\pi(w|k)$, and $\pi(\theta|w, k)$, respectively—the corresponding posterior in $\{k, w, \theta\}$ becomes

$$\pi(k, w, \theta|y) \propto \pi(k)\pi(w|k)\pi(\theta|w, k)L(y|\theta, w, k), \tag{38}$$

which we can integrate over $\{w, \theta\}$ to yield, $\pi(k|y)$. To this end we suppose a strictly finite mixture and run MC^3+RLR over a simple fixed grid of $k \in \{3, \ldots, 10\}$, estimating the series of (k-conditional) marginal likelihoods,

$$Z_k = \int \pi(w|k)\pi(\theta|w, k)L(y|\theta, w, k)dwd\theta, \tag{39}$$

to which we then apply $\pi(k)$ and normalize for $\pi(k|\theta)$; the computational details of this procedure are given in Section 4.3 below.

We now discuss a number of astronomical considerations relevant to our choice of priors for this particular modeling problem.

4.2 Astronomical Motivations for our Priors

As noted earlier, by considering the well-defined selection function of their observational campaign the authors of the original astronomical study were able to construct the expected probability density function of recession velocities for their survey under the null hypothesis (of a uniform distribution of galaxies throughout space). In particular, Postman et al. (1986) recognised that the strict *apparent magnitude* limit of their spectroscopic targetting strategy ($m_r < 15.7$ mag) would act as a luminosity (or *absolute magnitude*) limit evolving with distance according to

$$M_{r,\lim}(v) \approx m_r - 5\log_{10}(v) - 30, \tag{40}$$

where we have assumed v in units of 1000 km s^{-1} and a "Hubble constant" of $H_0 = 100$ km s^{-1} Mpc^{-1}. To estimate the form of the resulting selection function, $S_{\text{mag}}(v)$, Postman et al. (1986) considered how the relative number of galaxies per unit volume above (i.e., brighter than) this limit would vary with distance given the absolute magnitude distribution function, $F_{\text{mag}}(\cdot)$, for galaxies in the local Universe, i.e., $S_{\text{mag}}(v) \propto 1 - F_{\text{mag}}(M_{r,\lim}(v))$. To approximate the latter the astronomers simply integrated over a previous estimate of the local luminosity density parameterized as a Schechter function (Schechter, 1976) with characteristic magnitude, $M_r^* \approx -19.40 - 1.5$ mag, and faint-end slope, $\alpha_r^* \approx -1.3$, such that

$$f(M) \propto [10^{2/5(M_r^* - M)}]^{\alpha_r^* + 1} \exp[-10^{2/5(M_r^* - M)}], \tag{41}$$

and

$$S_{\text{mag}}(v) \propto \int_{-\infty}^{M_{r,\lim}(v)} f(M)dM. \tag{42}$$

An interesting feature of magnitude-limited astronomical surveys is that, although with increasing recession velocity this $S_{\text{mag}}(v)$ selection function restricts their sampling to the *decreasing* fraction of galaxies above $M_{r,\text{lim}}(v)$, the volume of the Universe probed by (the projection into three-dimensional space of) their two-dimensional angular viewing window is, in contrast, rapidly *increasing*. Hence, there exists an important additional selection effect, $S_{\text{vol}}(v)$, operating in competition with, and initially dominating, that on magnitude, and scaling with (roughly) the third power of distance, or recession velocity, such that

$$S_{\text{vol}}(v) \propto v^3. \tag{43}$$

The product of these two effects therefore returns the *net* selection function of the galaxy dataset, which we illustrate (along with each effect in isolation) in Figure 4 (see also Figure 4b from Postman et al. 1986). Important to note is that the net selection function is distinctly non-uniform; rather than lying flat across the range of v, it actually rises to a maximum of almost twice its initial power at $v \approx 17$, beyond which it declines slowly again towards zero at $v \approx 40$. As a result it is difficult to justify a strictly *symmetric* prior on the component weights, w, as has been popular in past analyses (e.g. Roeder & Wasserman 1997; Richardson & Green 1997). We therefore favor instead an *asymmetric* Dirichlet prior, $\pi(w|k) \sim \mathcal{D}(\delta)$, with weights, $\delta = \{\binom{k}{j}^{\gamma},\ j = 1, \ldots, k\}'$, and controlling hyperparameter, $\gamma > 0$, on the *mean-ordered* components (i.e., $\mu_1 < \ldots < \mu_k$). With the resulting Dirichlet density quite sensitive to the specified γ we have elected to treat this hyperparameter conservatively, specifying a nominal value of 0.2, but monitoring carefully the prior-sensitivity of $\pi(\theta|y)$ to this assumption in Section 4.4.

The selection function of the original survey can also meaningfully inform our prior on the distribution of component means (the $\mu_j \in \theta_j$). Here we adopt a common Normal prior across all μ_j, given in mean–precision form as $\pi(\mu_j) \sim \mathcal{N}(\kappa, \xi^{-1})$; with the hyperparameters, $\{\kappa = 17, \xi = 0.015\}$, chosen to give a reasonable match to the shape of $S_{\text{mag}}(v)S_{\text{vol}}(v)$. Interestingly, this κ is quite close to that of 20 chosen (*a posteriori*!) by Richardson & Green (1997) and others, though its corresponding ξ is significantly more informative than their choice of 0.0016 (which is so broad as to place ∼20% of the prior mass on components with *negative* mean recession velocities). For the precision of each mixture component (the $\tau_j \in \theta_j$) we adopt a common Gamma prior, $\pi(\tau_j) \sim \Gamma(\alpha, \beta)$; the hyperparameters of which are perhaps not so well constrained on astronomical grounds—although we can at least be confident that any large-scale clustering should occur above the scale of individual galaxy clusters (∼1 Mpc, or $\Delta v \approx 0.1$) and (unless the uniform space-filling hypothesis were correct) well below the width of our selection function. Hence, we simply adopt a fixed shape (hyper-)parameter of $\alpha = 2$ and allow the rate (hyper-)parameter to vary as a hyperprior with density, $\pi(\beta) \sim \Gamma(1, 0.05)$. Our choice here is again comparable to that of Richardson & Green (1997) who suppose $\pi(\beta) \sim \Gamma(0.2, 0.016)$—not $\Gamma(0.2, 0.573)$ as mis-quoted by Aitkin (2001)—though we evidently place far less prior weight on exceedingly large precisions (small variances).

Finally, to construct a prior for the number of components in the mixture we refer to yet another aspect of the Postman et al. (1986) survey design; namely, the deliberate placement of the five individual windows of the "unfilled" sample

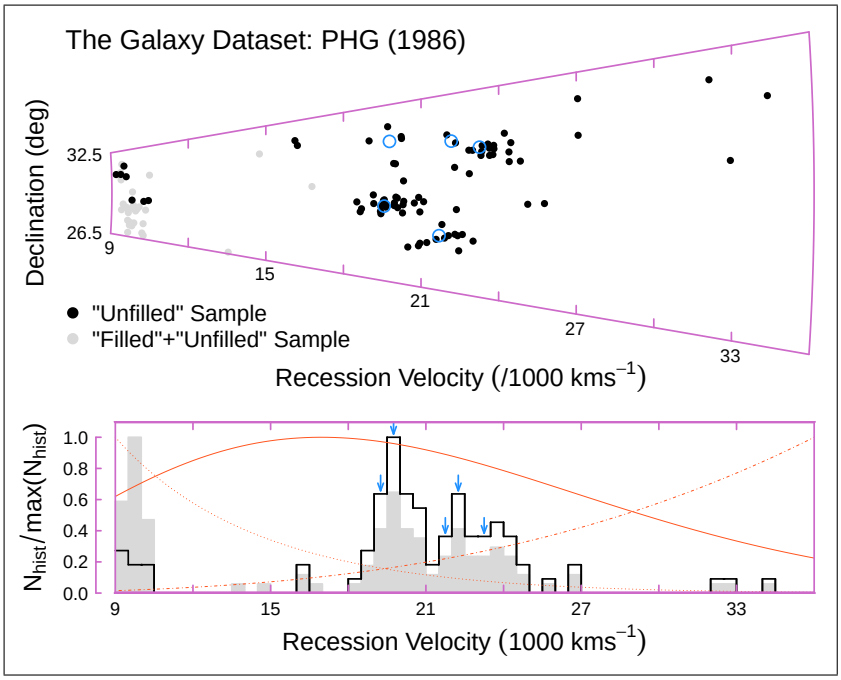

FIG 4. *Visualization of the galaxy dataset, including its Abell clusters and selection function. The clustering of galaxies in (two-dimensional) recession velocity–declination space is illustrated by way of the "cone diagram" shown in the top panel, and its (one-dimensional) projection to a recession velocity histogram shown in the bottom panel. In both cases the "unfilled" sample of the nominal galaxy dataset and the combined "filled"+"unfilled" sample of Postman et al. (1986) are indicated (with black and grey markings, respectively) for reference. The positions of five Abell clusters targetted by the "unfilled" survey are highlighted here in blue (circles/arrows), while the corresponding magnitude-dependent, volume-dependent, and net selection functions are overplotted (bottom panel) in red (the dotted, dash-dotted, and solid curves, respectively).*

over five previously-identified galaxy clusters from the Abell catalog (cf. their Figure 3), which anchors the mode of our Poisson $\pi(k)$, i.e., $\mathcal{P}(\lambda = 5)$. With the $k = 1$ and $k = 2$ mixture models already well excluded by previous analyses, and $k > 10$ a pragmatic upper bound for exploration given $\lambda = 5$, we explicitly truncate our prior to the domain, $3 \leq k \leq 10$. This contrasts somewhat with the Uniform priors on $k \leq 10$ and $k \leq 30$ assumed by Roeder & Wasserman (1997) and Richardson & Green (1997), respectively—though reweighting for alternative $\pi(k)$ is trivial even without RLR.

4.3 Application of RLR with MC^3 Sampling

As Lee et al. (2008) highlight in their review of contemporary methods for inference on Bayesian mixture models, whether for Gibbs sampling (e.g. Diebolt & Robert 1994), random walk MCMC, or otherwise, efficient exploration of the mixture model posterior at fixed k can be difficult to achieve due to the intrinsic non-identifiability of the component means (not present here with our asymmetric $\pi(w_i)$) and the imposition of a Dirichlet prior structure on the component weights (*viz.* the simplex constraint, $\sum_{j=1}^{k} w_j = 1$ with $0 < w_j < 1$). Here we have opted for an MC^3 strategy to accomplish this task, taking advantage of the RLR requirement for exploring multiple bridging densities to hasten the convergence of our tempered-likelihood MCMC chains. Rather than update via Gibbs sampling,

which can become 'stuck' on particular permutations of the allocation vector, z, we follow a simple random-walk Metropolis-Hastings prescription.

For the present analysis (conducted on a modest laptop computer) we run our MC^3 sampler for each k with twenty temperatures spaced uniformly as $t = \{(0, 1/19, \ldots, 19/19)\}$ (giving a reasonably equal spacing of the log $\hat{Z}_{t(k')}$ for this problem), a burn-in phase of 10,000 draws, and a final output of 100,000 draws from each $t(k')$. We then thin each output chain to 5000 draws (for computational speed) and run RLR with Kong et al. (2003)'s form for the asymptotic covariance matrix to estimate our uncertainties on the derived log $\hat{Z}_k$. The results of this analysis are illustrated in Figure 5. Under the astronomically-motivated priors stipulated in Section 4.2 our marginal likelihood computations clearly favor a six-to-eight component mixture; the form of which we illustrate for the $k = 7$ case (in comparison to a histogram of the galaxy dataset) in the right-hand panel of Figure 5. A quick inspection of the latter, however, suggests that three of the seven mixture components are being used here to model what might reasonably be considered a single, or perhaps double, component of coherent velocity structure. Furthermore, had we adopted a flat $\pi(k)$ instead of a $\lambda = 5$ (truncated) Poisson the resulting posterior would have peaked towards even higher k (eight-to-nine), suggesting an even greater degree of 'over-fitting'.

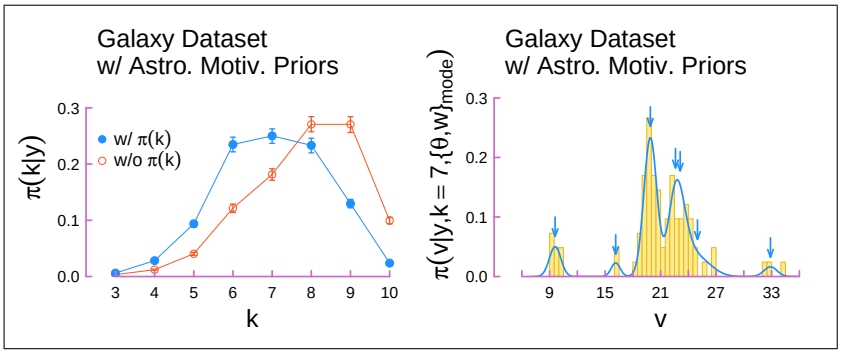

FIG 5. *Posterior probabilities for the number of Normal mixture components in the galaxy dataset, $\pi(k|y)$, under our astronomically-motivated priors (left-hand panel). The solid, blue symbols here denote the true posterior, while the open, red symbols indicate for reference the raw marginal likelihood-based result, i.e., before application of our Poisson $\pi(k)$. In each case the relevant uncertainties (recovered from estimates of the asymptotic covariance matrix for each* log $\hat{Z}_k$*) are illustrated as 1σ error bars. The inferred probability density (in velocity space) at the maximum a posteriori parameterization of our Normal mixture model for $k = 7$ is then illustrated for reference against a scaled histogram of the galaxy dataset in the right-hand panel.*

Given the lack of any strong physical motivation for the assumed Normal form, a significant non-Normality of the underlying components could well be a good explanation for this result; and indeed such a hypothesis might readily be investigated by extending the above model comparison across mixtures of Student's t or skew Normal distributions as well. However, within the Normal mixture model framework to which we confine ourselves for the present example another explanation could lie in the sensitivity of the posterior to the chosen priors, and so we investigate this with our RLR output in Section 4.4 below.

4.4 Exploration of Prior-Sensitivity

As noted in Section 1.4, by estimating normalization constants not just for the posterior itself but also for the specified sequence of bridging densities linked with the prior (or suitable auxiliary) the RLR framework readily faciliates the analysis of prior-sensitivity via importance-sampling-based reweighting of the output draws (cf. Equation 29). We therefore begin here by confirming the flexibility of this reweighting procedure, using our RLR output to recover $\pi(k|y)$ under the priors used by Richardson & Green (1997) while keeping track of our uncertainties via the effective sample size (as per Kong et al. 2003); the success of which is illustrated in the left-hand panel of Figure 6. At face value the greatest difference between our prior construction and that of Richardson & Green (1997) might be considered the asymmetry and symmetry, respectively, of the Dirichlet distributions chosen for our component weights. As explained in Section 4.2 the asymmetry of our $\pi(w)$ is controlled by the parameter, γ, which we set as default to a value of 0.2. To investigate the sensitivity of $\pi(k|y)$ to this parameter we therefore reweight with two alternatives, $\gamma = 0$ (symmetric) and $\gamma = 0.4$ (more asymmetric), keeping all other hyperparameters fixed; the results of which are shown in the right-hand panel of Figure 6. Interestingly, we find that our default choice of $\gamma = 0.2$ has had only a minor role in shaping the posterior, relative to the symmetric alternative, although a γ as high as 0.4 does indeed force $\pi(k|y)$ to even further 'over-fitting'.

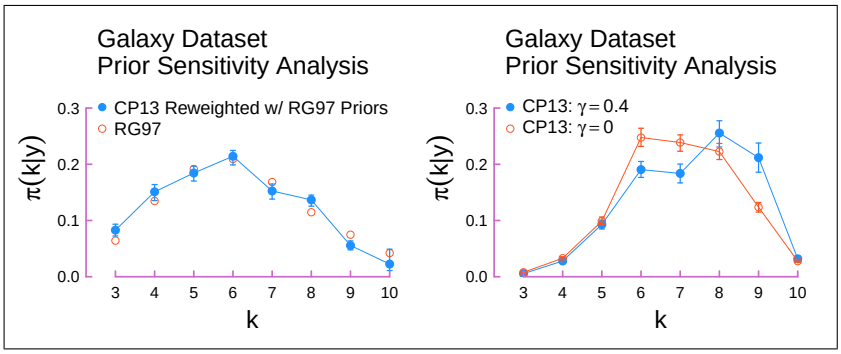

FIG 6. *Demonstration of the potential for reweighting of RLR output to investigate the sensitivity of Bayesian model selection to the specified priors. In the left-hand panel we reweight our draws from the galaxy dataset posterior to match the priors used by Richardson & Green (1997), keeping track of our uncertainties via the effective sample size. In the right-hand panel we investigate the effect of modifying the parameter, γ, controlling the asymmetry of our Dirichlet prior, revealing that this factor alone is insufficient to account for the marked differences between our $\pi(k|y)$ and that of Richardson & Green (1997).*

Continuing with our prior-sensitivity analysis we have recomputed $\pi(k|y)$ under two and four-fold increase and decreases in each of our remaining hyperparameters. Of these the most marked sensitivity was observed for the precision, ξ, of our Normal prior on the *positions*, μ_j, of the component means; recalling that $\xi = 0.015$ was chosen to match the shape of the selection function (in velocity space) of the original astronomical survey. As we illustrate in the left-hand panel of Figure 7 a factor of four decrease to $\xi = 0.00375$ is in fact sufficient to account for much of the difference between our nominal posterior and that of Richardson & Green (1997), shifting our $\pi(k|y)$ towards significantly fewer mix-

ture components. Moreover, as we show in the right-hand panel of this Figure the $k = 6$ model favored under this modified prior appears visually to give a far more satisfactory fit to the binned data. This is of course a little ironic as such a broad prior on the component means is much less appealing from an astronomical perspective (in that it places significant probability on detecting components far outside the bounds of the true selection function). Nevertheless, we hope that the analysis presented here has given a clear demonstration of the potential of RLR-based marginal likelihood computation for exploring prior-sensitivity in the Bayesian model selection context.

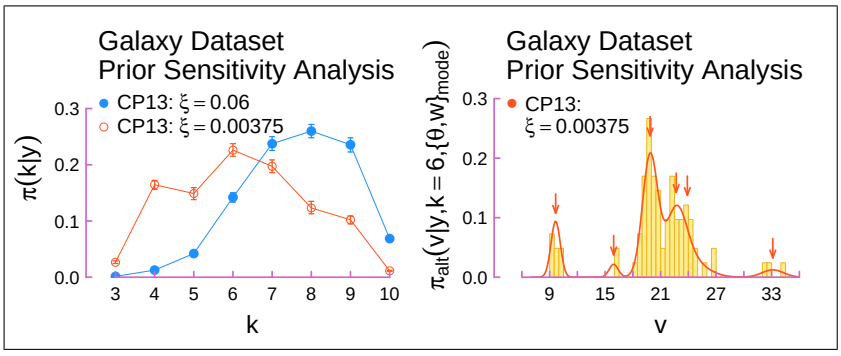

FIG 7. *Continuing demonstration of the potential for reweighting of RLR output to investigate the sensitivity of Bayesian model selection to the specified priors. In the left-hand panel we investigate the effect of modifying the parameter, ξ, controlling the precision of our Normal prior on the position of the component means. Somewhat surprisingly (see discussion) we find $\pi(k|y)$ for the galaxy dataset markedly responsive to modest changes in this particular hyper-parameter; with a factor of four decrease in our nominal $\xi = 0.015$ alone almost sufficient to match the Richardson & Green (1997) result. The inferred probability density (in velocity space) at the maximum a posteriori parameterization of our Normal mixture model under this rescaled hyper-parameter for $k = 6$ is illustrated for reference against a scaled histogram of the galaxy dataset in the right-hand panel.*

5. CONCLUSIONS

In this paper we have explored both the theoretical foundations of and connections between those recursive pathways to marginal likelihood estimation characterized by reverse logistic regression and the density of states, and we have introduced the novel heuristic of 'thermodynamic integration via importance sampling' for better understanding the role of the bridging sequence in this process. Furthermore, by way of our numerical examples we have highlighted a number of considerations for maximizing the efficiency of RLR-type schemes (tailoring of the bridging sequence to achieve an equal spacing of $\log \hat{Z}_k$; use of a data-driven reference/auxiliary; use of all draws in nested ellipse-based sampling), and, importantly, their utility for prior-sensitivity analysis. Though a 'one-size-fits-all' algorithm to solve the challenging problem of marginal likelihood estimation remains elusive we hope that our contribution herein leads to both greater use of the recursive pathway itself and greater interest in estimators that facilitate rapid recomputation under alternative priors in general.

REFERENCES

Aitkin, M. (2001). Likelihood and Bayesian analyisis of mixtures. *Statist. Mod.* **1** 287–304.

Arima, S., Tardella, L. (2012). Improved harmonic mean estimator for phylogenetic model evidence. *J. Comput. Biol.* **19** 418–438.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29** 2157–2167.

Calderhead, B., Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Statist. Data Anal.* **53** 4028–4045.

Caimo, A., Friel, N. (2012). Bayesian model selection for exponential random graph models. [*preprint*] arXiv:1201.2337

Cappé, O., Guillin, M., Marin, J. M., Robert, C. P. (2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13** 907–930. MR2109057

Chen, M.-H., Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25** 1563–1594. MR1463565

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

Chopin, N., Robert, C. P. (2010). Properties of nested sampling. *Biometrika* **97** 741–755. MR2672495

Davis, T. M., et al. (2007). Scrutinizing exotic cosmological models using ESSENCE supernova data combined with other cosmological probes. *Astroph. J.* **666** 716–725.

Del Moral, P., Doucet, A., Jasra, A. (2006). Sequential Monte Carlo samplers. *J. Roy. Statist. Soc. B* **68** 411–436. MR2278333

Diebolt, J., Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56** 363–375. MR1281940

Escobar, M. D., West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

Evans, M., Robert, C. P., Davison, A. C., Jiang, W., Tanner, M. A., Doss, H., Qin, J., Fokianos, K., MacEachern, S. N., Peruggia, M., Guha, S., Chib, S., Ritov, Y., Robins, J. M., Vardi, Y. (2003). Discussion on the paper by Kong, McCullagh, Meng, Nicolas and Tan. *J. Roy. Statist. Soc. B* **65** 604–618.

Fan, Y., Rui, W., Chen, M.-H., Kuo, L., Lewis, P. O. (2012). Choosing among partition models in Bayesian phylogenetics. *Mol. Boil. Evol.* **28** 523–532.

Feroz, F., Hobson, M. P. (2008). Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Mon. Not. R. Astron. Soc.* **384** 449–463.

Ferrenberg, A. M., Swendsen, R. H. (1989). Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63** 1195–1198.

Friel, N., Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. Roy. Statist. Soc. B* **70** 589–607. MR2420416

Friel, N., Wyse, J. (2012). Estimating the evidence–a review. *Stat. Neerl.* **66** 288–308.

Gelfand, A. E., Dey, D. (1994). Bayesian model choice: asymptotic and exact calculations. *J. Roy. Statist. Soc. B* **56** 501–514. MR1278223

Gelman, A., Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507

Gelman, A. B., Carlin, J. B., Stern, H. S., Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton. MR2027492

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

Geyer, C. J., Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. B* **54** 657–699. MR1185217

Geyer, C. J. (1994). *Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo*, Technical Report 568. School of Statistics, University of Minnesota, Minneapolis.

Gunn, J. E., Gott, J. R. III (1972). On the infall of matter into clusters of galaxies and some effects on their evolution. *Astrophys. J* **176** 1–19.

Habeck, M. (2012). Evaluation of marginal likelihoods via the density of states. *J. Mach. Learn. Res.* (*Proceedings Track*) **22** 486–494.

Hesterberg, T. (1991). Weighted average importance sampling and defensive mixture distributions. *Technical Report 148*. Stanford University, Stanford.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors). *Statist. Sci.* **14** 382–417. MR1765176

Hörmander, L. (1983). *The analysis of linear partial differential operators. I. Distribution theory and Fourier analysis.*, Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin. MR0717035

Hubble, E. (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Proc. Natl. Acad. Sci. USA* **15** 168–173.

Jaynes, E. T. (2003). *Probability theory: the logic of science.* Cambridge University Press, Cambridge. MR1992316

Jeffreys, H. (1961). *Theory of probability.* 3rd ed. Clarendon Press, Oxford. MR0187257

Jeffreys, W. H., Berger, J. O. (1991). *Sharpening Ockham's razor on a Bayesian strop.* Technical Report #91-44C. Department of Statistics, Purdue University, Indiana.

Kass, R. E., Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

Kilbinger, M., Wraith, D., Robert, C. P., Benabed, K., Cappé, O., Cardoso, J.-F., Fort, G., Prunet, S., Bouchet, F. R. (2010). Bayesian model comparison in cosmology with population Monte Carlo. *Mon. Not. R. Astron. Soc.* **405** 2381–2390.

Kirshner, R. P., Oemler, A. Jr., Schechter, P. L., Shectman, S. A. (1981). A million cubic megaparsec void in Boötes? *Astrophys. J* **248** 57–60.

Kong, A., Liu, J. S., Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., Tan, Z. (2003). A theory of statistical models of Monte Carlo integration. *J. Roy. Statist. Soc. B* **65** 585–618. MR1998624

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13** 1011–1021.

Lartillot, N., Phillipe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.* **55** 195–207.

Lee, K., Marin, J.-M., Mengersen, K., Robert, C. P. (2008). Bayesian inference on mixtures of distributions. [*preprint*] arXiv:0804.2413

Lefebvre, G., Steele, R., Vandal, A. C. (2010). A path sampling identity for computing the Kullback-Leibler and J divergences. *Comput. Statist. Data Anal.* **54** 1719–1731. MR2608968

Li, Y., Ni, Z.-X., Lin, J.-G. (2011). A stochastic simulation approach to model selection for stochastic volatility models. *Comm. Statist. Simulation Comput.* **40** 1043–1056. MR2792481

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics. Springer-Verlag, New York. MR1842342

MacLachlan, G., Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Statistics: Applied Probability and Statistics. Wiley-Interscience, New York. MR1789474

Meng, X.-L., Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

Mukherjee, P., Parkinson, D., Liddle, A. R. (2006). A nested sampling algorithm for cosmological model selection. *Astrophys. J.* **638** L51–L54.

Neal, R. (1999). Erroneous results in "Marginal likelihood from the Gibbs output". URL(http://www.cs.toronto.edu/∼radford/ftp/chib-letter.pdf)

Neal, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. MR1837132

Newton, M. A., Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. B* **56** 3–48. MR1257793

Phillips, D. B., Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte-Carlo in Practise*, W.R.Gilks, S.T. Richardson and D.J. Spiegelhalter (Eds.) 215–240.

Postman, M., Huchra, J. P., Geller, M. J. (1986). Probes of large-scale structure in the Corona Borealis region. *Astrophys. J.* **92** 1238–1247.

Raftery, A. E., Newton, M. A., Satagopan, J. M., Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics* **8** 371–416. MR2433201

Richardson, S., Green, P. J. (2007). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B* **59** 731–792. MR1483213

Robert, C. P., Wraith, D. (2009) Computational methods for Bayesian model choice. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: The 29th International*

Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. (*AIP Conference Proceedings*), vol. 1193, pp. 251–262.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Am. Statist. Assoc.* **85** 617–624.

Roeder, K., Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121

Schechter, P. (1976). An analytic expression for the luminosity function of galaxies. *Astrophys. J.* **203** 297–306.

Shirts, M. R., Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129** 124105.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Anal.* **1** 833–860. MR2282208

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903

Tan, Z., Gallicchio, E., Lapelosa, M., Levy, R. M. (2012). Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J. Chem. Phys.* **136** 144102.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–203. MR773161

Wolpert, R. L., Schmidler, S. C. (2012). α-Stable limit laws for harmonic mean estimators of marginal likelihoods. *Statist. Sinica* **22** 1233–1251.

Xie, W., Lewis, P., Fan, Y., Kuo, L., Chen, M.-H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **18** 1001–1013.